Advances in Applied Mathematics 应用数学进展, 2022, 11(8), 5936-5950
Published Online August 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.118625

Hans汉斯

线性回归方法在空气质量影响因素分析中的应用

刘一鹤，牟唯嫣，金　童

北京建筑大学理学院，北京

收稿日期：2022年7月22日；录用日期：2022年8月15日；发布日期：2022年8月25日

摘　要

近年来，随着社会经济发展和人们生活水平的提高，环境污染越来越受到人们的重视。在评估某地区环境好坏时，空气质量是其中一个重要的指标。同时，空气质量也与每个人息息相关。为此，本文通过2020年全国各省市的空气质量数据及各省相应经济数据、人口数据、公共预算支出等数据，使用最小二乘法、逐步回归法、岭回归法、Lasso回归、主成分回归法五种方法构建了多元线性回归模型。利用得到的模型进行分析，结果表明：人均可支配收入、人口数、工业产品产量与空气质量呈负相关；公共预算支出中的科学技术预算支出与空气质量呈正相关。并且工业产品产量及科学技术预算支出对空气质量的影响较为显著。除此之外，本文对比了五种线性回归方法建模精度，得到在存在多重共线性时，逐步回归法与主成分法预测精度较高。其余四种方法都明显好于普通最小二乘法。

关键词

线性回归，空气质量，多重共线性

Application of Linear Regression Methods in the Analysis of Air Quality Influencing Factors

Yihe Liu, Weiyan Mu, Tong Jin

School of Science, Beijing University of Civil Engineering and Architecture, Beijing

Received: Jul. 22nd, 2022; accepted: Aug. 15th, 2022; published: Aug. 25th, 2022

Abstract

In recent years, with the development of social economy and the improvement of people's liv-

文章引用: 刘一鹤, 牟唯嫣, 金童. 线性回归方法在空气质量影响因素分析中的应用[J]. 应用数学进展, 2022, 11(8): 5936-5950. DOI: 10.12677/aam.2022.118625

ing standard, more and more people pay attention to environmental pollution. Air quality is an important index in evaluating the environment of a certain area. At the same time, air quality is closely related to everyone. Therefore, based on the air quality data of all provinces and cities in 2020 and the corresponding economic data, population data and public budget expenditure data of each province, this paper constructed a multiple linear regression model using five methods including least square method, stepwise regression method, mountain regression method, Lasso regression method and principal component regression method. The results show that per capita disposable income, population and industrial product output are negatively correlated with air quality. Science and technology budget expenditure in public budget expenditure is positively correlated with air quality. The output of industrial products and budget expenditure of science and technology have a significant impact on air quality. In addition, this paper compares the modeling accuracy of five linear regression methods, and obtains that the prediction accuracy of stepwise regression method and principal component method is higher in the presence of multicollinearity. The other four methods are obviously better than ordinary least square method.

Keywords

Linear Regression, Air Quality, Multicollinearity

1. 引言

随着我国经济的不断发展，党的十九大指出，我国社会的主要矛盾转化为人民日益增长的美好生活需要同不平衡不充分的发展之间的矛盾。习近平主席说过：“绿水青山就是金山银山”。因此，从国家与社会、个人层面，都对环境质量越来越重视。而环境保护中的空气质量与每个人都息息相关。为探讨各个指标对空气质量的影响，黄煜宁等人以福州市 PDPI 作为解释变量，空气综合污染指数作为被解释变量，得出 2004~2018 年期间，福州环境空气质量与经济增长之间的关系[1]。金仁浩等人在对各区域空气质量数据进行描述分析的基础上，从整体上分析北京市空气质量与气象、社会经济因素之间的关系[2]。刘亦文通过选取中国 30 个重点城市 2003~2018 年的面板数据，考察碳排放总量和强度双约束政策对城市空气质量的平均因果效应[3]。基于此，本文通过各省空气质量指标与各省经济指标、工业产品产量等数据建立多元线性回归模型，探讨哪些因素对空气质量有影响。为改善空气质量、保护环境应进行哪些举措提供建议。同时探讨各种回归方法，在处理具有多重共线性的数据时的表现。

本文第一部分介绍了对空气质量影响因素的研究；第二部分对数据进行初步处理与分析；第三部分使用最小二乘法、岭回归、Lasso 回归、主成分回归法对数据进行建模，得到相应的模型及参数。第四部分对模型精度进行比较，对各个模型的参数进行解释。通过分析得到的模型及参数，得出环境污染与工业产量、经济发展、政府预算的相关关系。

2. 数据录入及初步处理

原始数据使用全国各省会城市空气质量及达到二级的天数作为因变量 Y，其中自变量 M 代表各省市

人均可支配收入；自变量 G 代表每省市人均生产总值；自变量 P 为各省人口数；自变量 B 为各省市当年环境保护预算支出；自变量 S 为各省当年科技预算支出；自变量 I 是各省当年部分工业产品产量。工业产品产量合计中，选择了部分对环境污染较高且数值较大的工业产品。上述数据均为 2020 年数据，来自《中国统计年鉴 2021》。原始数据见附表 1。

由于自变量的量纲不同，为了使数据具有参考性，首先对数据进行标准化。利用 R 软件中的 scale 函数，可以得到原始数据进行中心化标准化后的结果。数据标准化结果见附表 2。

为了检验标准化后的数据是否具有多重共线性，使用三种方法。利用 cor 函数、vif 函数，分别计算数据的相关系数、VIF 值及条件数。得到结果见表 1、表 2。其中表 1 保留四位小数。

Table 1. Correlation coefficient
表 1. 相关系数

	Y	*M*	*G*	*P*	*B*	*I*	*S*
Y	1.0000	−0.0460	−0.0213	−0.2789	−0.3997	−0.5846	0.0051
M	−0.0460	1.0000	0.9539	0.0478	0.1981	−0.1672	0.5840
G	−0.0213	0.9539	1.0000	0.0565	0.1874	−0.1478	0.5857
P	−0.2789	0.0478	0.0565	1.0000	0.8108	0.3241	0.6841
B	−0.3997	0.1981	0.1874	0.8103	1.0000	0.4739	0.6619
I	−0.5846	−0.1672	−0.1478	0.3241	0.4739	1.0000	0.0040
S	0.0051	0.5840	0.5857	0.6841	0.6619	0.0040	1.0000

Table 2. Condition number and VIF value
表 2. 条件数及 VIF 值

条件数	自变量	*M*	*G*	*P*	*B*	*I*	*S*
70.54245	VIF 值	11.889480	11.606148	4.693508	4.280061	1.684187	5.160048

由表 1 可知，有些自变量间相关系数大于 0.5，例如：自变量 M 与自变量 G 的相关系数约为 0.95。因此，可以认为这些系数之间具有相关性。数据具有多重共线性。由表 2 可得，条件数小于 100，可以认为共线程度较小；但自变量 M 与自变量 G 的 VIF 值均大于 10，认为数据具有多重共线性。综合三种方法，认为我们所选数据具有多重共线性。因此，在后文中选择的岭回归法、Lasso 回归法、主成分法都可以部分的消除数据多重共线性，使得到的结果更加准确。

接下来检验数据是否服从正态分布。本文使用了密度图与 QQ 图，直观检验数据是否服从或近似服从正态分布。除此之外，使用 Shapiro-Wilk 方法，从数值上准确判断数据是否通过正态性检验。密度图与 QQ 图见图 1、图 2。

通过图 1，可以看到标准化后数据的曲线近似为钟形；从图 2 得到数据的样本点基本分布在 45 度参考线内。通过观察，可以认为数据近似服从正态分布。再利用 R 软件中 shapiro.test 函数，得到对数据进行 Shapiro-Wilk 检验的 p 值为：0.08118 > 0.05。综上，我们认为数据基本符合正态分布。因此，我们可以使用标准化后的数据进行下面的统计分析。

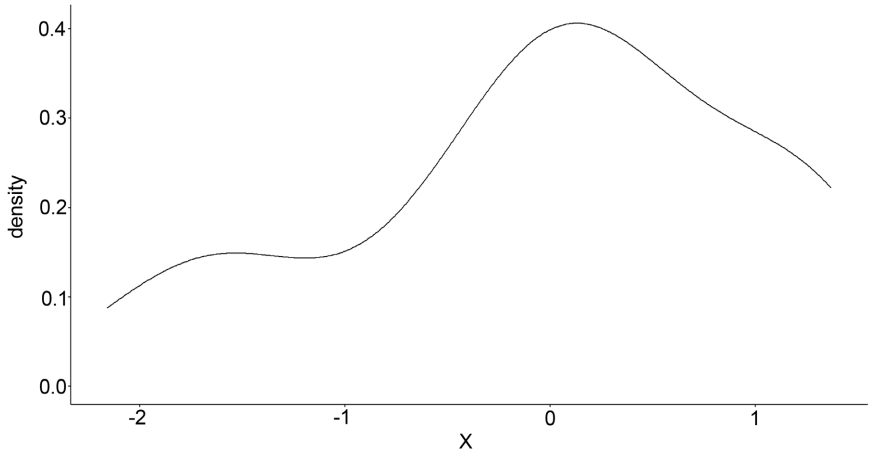

Figure 1. Density graph
图 1. 密度图

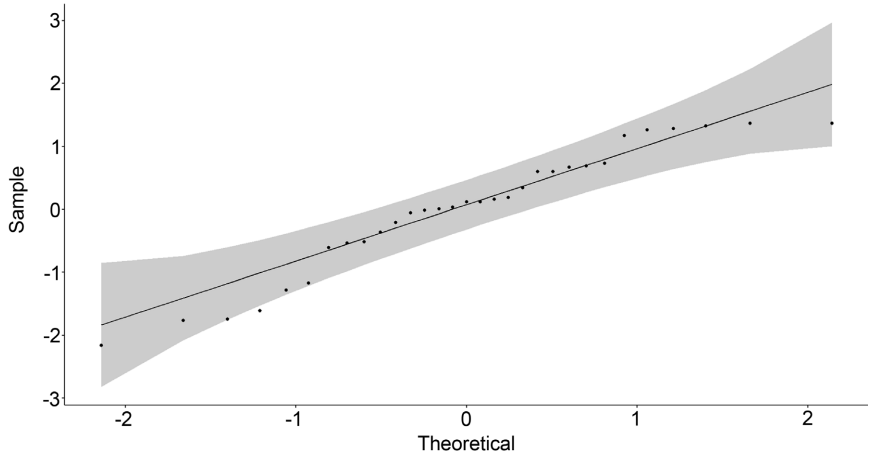

Figure 2. QQ Chart
图 2. QQ 图

3. 最小二乘法

3.1. 普通最小二乘法及逐步回归法介绍

3.1.1. 最小二乘法介绍

若有回归方程记为：

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon$$

其中 $\beta_1, \beta_2, \cdots, \beta_m$ 为回归系数，也是待估参数。ε 是随机向量。若 $\hat{\beta}_i$ 是参数估计量，y_k 是观测值。那么最小二乘法满足：

$$Q = \sum_{k=1}^{n} \left[y_k - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m \right) \right]^2$$

达到最小，就可以解出 $\hat{\beta}$。$\hat{\beta}$ 就是 β 的最小二乘估计。

3.1.2. 逐步回归法介绍

逐步回归法是利用最小二乘法原理，通过建立正规方程，来解回归系数的一种回归方法。他是向前

选择变量法和向后剔除变量法的组合，每进行一步只从中选取一个变量，能够使得到的回归方程更加简化，变量之间的相关性增大[4]。

3.2. 普通最小二乘法建模

尽管经过检验，本例不适合使用普通最小二乘法，但仍使用最小二乘法进行分析。将普通最小二乘的结果作为之后各种方法的参考，与之进行比较。

利用 R 软件中 lm 函数，可以得到模型为：

$$Y = -0.58M + 0.11G - 0.42P - 0.26B - 0.41I + 0.74S$$

使用普通最小二乘法得到的描述性统计量见表 3。

Table 3. Descriptive statistics for ordinary least squares regression
表 3. 普通最小二乘回归的描述性统计量

自变量	系数估计值	标准误差	t 值	p 值
截距	0.0000	0.1436	0.0000	1.0000
M	−0.5845	0.5032	−1.1620	0.2568
G	0.1139	0.4972	0.2290	0.8207
P	−0.4201	0.3162	−1.3290	0.1964
B	−0.2632	0.3019	−0.8720	0.3920
I	−0.4076	0.1894	−2.1520	0.0416*
S	0.7428	0.3315	2.2410	0.0346*

*代表 p 值在 0.01 到 0.05 范围内，**代表 p 值在 0.001 到 0.01 范围内，***代表 p 值小于 0.001，下同。

由表 3 可知，只有变量 *I* 与变量 *S* 的 p 值小于 0.05，是显著的。因此，由普通最小二乘法得到的模型为：

$$Y = -0.41I + 0.74S$$

由于数据进行了标准化，故将系数还原后得到：

$$Y = 292.9536 - 0.0005161216I + 0.1612487S$$

3.3. 逐步回归法建模

在本例中，初始模型包含所有变量。根据 AIC 值选出最优模型，做逐步回归。利用 lm 函数与 step 函数，可以得到逐步回归法建立的模型为：

$$Y = -0.5M - 0.57P - 0.49I + 0.69S$$

此模型相应的描述性统计量见表 4。

Table 4. Descriptive statistics for the stepwise regression method of model construction
表 4. 逐步回归法建造模型的描述性统计量

自变量	系数估计值	标准误差	t 值	p 值
截距	0.0000	0.1404	0.0000	1.0000
M	−0.5036	0.2192	−2.2980	0.02987*
P	−0.5694	0.2657	−2.1430	0.04163*
I	−0.4870	0.1594	−3.0550	0.00515**
S	0.6906	0.3103	2.2260	0.03491*

各个自变量均显著，其 p 值均小于 0.05，因此最终模型即为上式。将标准化后的数据还原，得到的模型为：

$$Y = 394.8733 - 0.001802806M - 0.008439665P - 0.0006168282I + 0.1503535S$$

4. 岭回归

4.1. 岭回归介绍

岭回归估计是改进最小二乘估计的一种算法，可以解决最小二乘法在求解系数向量时矩阵无法求逆的缺点[5]。岭回归估计放弃了无偏性和部分精确度，可以得到效果稍差但更符合实际的结果。岭回归非常灵活，使用时存在一定的主观性，但这种主观人为的性质正好可以使得定性分析与定量分析进行有机结合，在解决多重共线性问题时有特殊的作用[6]。应用岭回归方法时，通常要使用标准化后的数据进行计算。首先用普通最小二乘法可以得到正规方程为[7]：

$$r_{xx}b = r_{yX}$$

其中，r_{xx} 是 X 的相关系数矩阵，r_{yX} 是 y 与 X 的相关系数的向量。因此，在岭回归中可以使用 $k(k \geq 0)$ 作为估计量的偏差，得到岭回归的正规方程是：

$$(r_{xx} + kI)b_R = r_{yX}$$

k 代表了岭回归中估计量的偏差值，这也是岭回归的特点。若 $k = 0$，上式为普通最小二乘回归方法；若 $k > 0$，则是岭回归方法，并且使得估计量 b_R 的均方误差小于最小二乘回归估计量 b 的均方误差。如果选取了理想的 k 值，就可以使估计量的偏差与方差组合后达到最好效果。因此，在岭回归中选择合适的 k 值来进行计算是非常重要的。

4.2. 岭回归建模

在 R 软件中加载 MASS 包，通过 lm.ridge 函数对标准化后的数据做岭回归。在选取 λ 值时，首先在 0~5 范围内选取，精确到 0.1，运算得到最优 λ 值为 4.6。再设置范围为 4~5，精确到 0.01，得到最优 λ 值为 4.65。因此，选择 λ 值为 4.65。利用图片直观确认 λ 值时，可以使用岭迹图，并且做出 λ 值与残差平方和的关系图。如图 3 所示，从图 3 中左图得到，回归系数都相对稳定；由图 3 的右图得到在 λ 取值为 4.65 时，残差平方和相对较小。

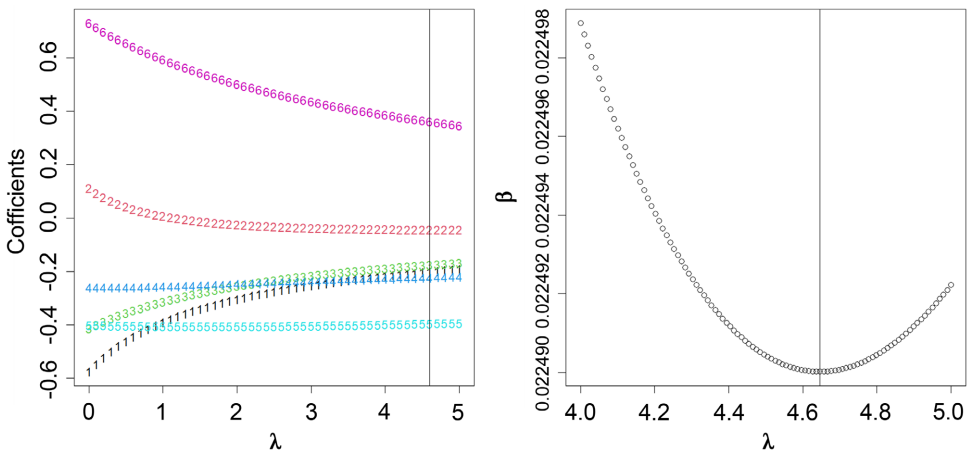

Figure 3. Ridge trace plot and λ-value versus residual sum of squares
图 3. 岭迹图及 λ 值与残差平方和关系图

确定 $\lambda = 4.65$ 后，使用岭回归方法得到的模型为：

$$Y = -0.01M - 0.04P - 0.06B - 0.1I + 0.01S$$

岭回归模型的描述性统计量见表 5。

Table 5. Descriptive statistics for the ridge regression method of model construction
表 5. 岭回归法建造模型的描述性统计量

自变量	系数估计值	标准误差	t 值	p 值
M	−0.0094	0.1355	0.3810	0.7033
G	−0.0038	0.1357	0.1540	0.8777
P	−0.0370	0.1388	1.4580	0.1448
B	−0.0585	0.1358	2.3580	0.018385*
I	−0.0968	0.1564	3.3910	0.000697***
S	0.0137	0.1326	0.5640	0.5726

由表 5 得只有变量 B 与变量 I 显著，因此使用岭回归方法对数据进行拟合的模型为：

$$Y = -0.06B - 0.1I$$

岭回归使用标准化后数据进行建模，故还原自变量的系数为：

$$Y = 312.9246 - 0.02380766B - 0.0001258833I$$

5. Lasso 回归

5.1. Lasso 回归介绍

Lasso 回归是一种“降维”思想的方法。在进行 Lasso 回归时，通过构造一个惩罚函数，可以减小变量的系数甚至将系数降为 0，这是岭回归无法做到的。Lasso 回归可以适用于线性与非线性两种情况，其原理是在普通的线性模型后，增加惩罚项 L1，通过惩罚项来调节系数，降低维度。

假设因变量为：$y = (y_1, y_2, \cdots, y_n)^{\mathrm{T}}$，自变量为：$X = (x_{1j}, x_{2j}, \cdots, x_{nj})^{\mathrm{T}}$，$j = 1, 2, \cdots, p$，$\beta = (\beta_1, \beta_2, \cdots, \beta_n)^{\mathrm{T}}$ 是系数向量，考虑线性模型[8]：

$$y = \beta X + \varepsilon$$

Lasso 方法的变量选择和参数估计可以由下式得到，其中 λ 是正则化参数：

$$\beta(\text{Lasso}) = \arg\min \sum_{i=0}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} |\beta_j|$$

在求解上式时，可以转化为带有惩罚项的优化问题，其中 t 与 λ 相对应，是调整参数：

$$\beta(\text{Lasso}) = \arg\min \sum_{i=0}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2 \quad \text{s.t.} \sum_{j=0}^{p} |\beta_j| \le t$$

5.2. Lasso 回归建模

由于 Lasso 回归要求数据为矩阵形式，因此设 x 与 y 为数据标准化后的自变量与因变量数据的矩阵。使用 glmnet 函数进行回归拟合，并通过交叉验证选择 λ 值，得到图 4、图 5。图 4 展示了随着 λ 值的变

化，变量系数的变化。图 5 中两条虚线代表了均方误差最小时的 λ 值及距离均方误差最小值一个标准误时得 λ 值。

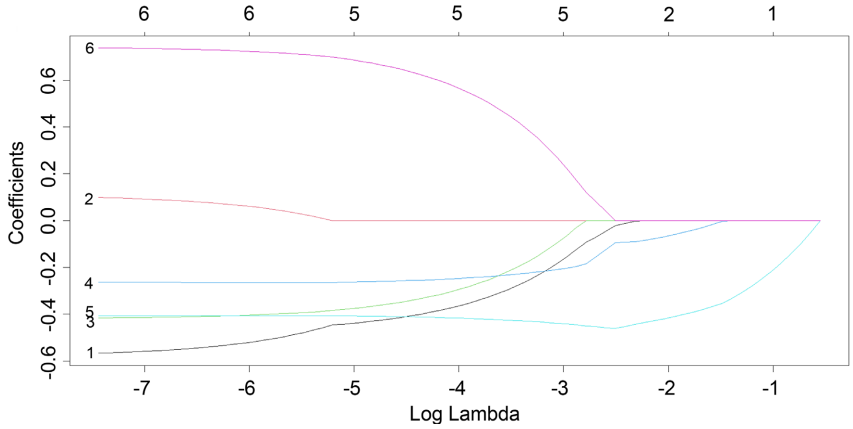

Figure 4. Variation of λ values and coefficients of variables
图 4. λ 值与变量系数的变化

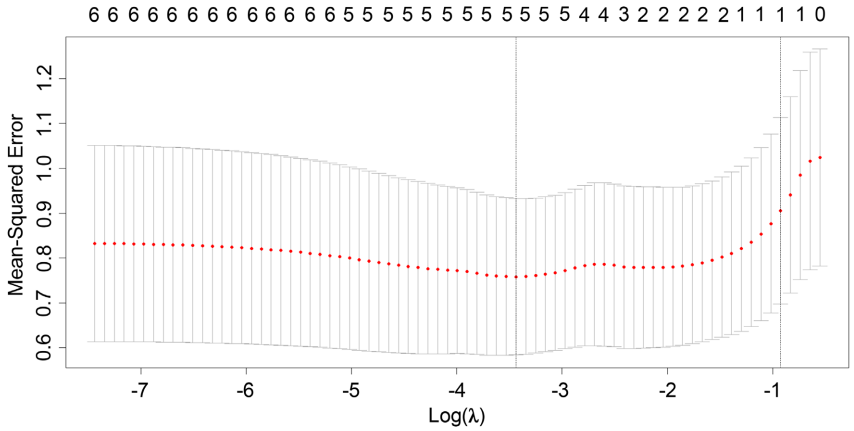

Figure 5. λ value and mean square error
图 5. λ 值与均方误差

由图及计算可得，在 Lasso 回归中可取：$\lambda_1 = 0.0322$，$\lambda_2 = 0.3964$。在 $\lambda = 0.3964$ 时，模型含有五个变量，因此选择 λ_2 作为 λ。得到的模型为：

$$Y = -0.48M - 0.42P - 0.27B - 0.4I + 0.76S$$

模型的描述性统计量见表 6。

Table 6. Descriptive statistics for the Lasso regression method of model construction
表 6. Lasso 回归法建造模型的描述性统计量

自变量	系数估计值	标准误差	t 值	p 值
截距	0.0000	0.1408	0.0000	1.0000
M	−0.4814	0.2211	−2.1770	0.0391*
P	−0.4231	0.3098	−1.3650	0.1843
B	−0.2718	0.2938	−0.9250	0.3637

Continued

I	−0.4022	0.1843	−2.1820	0.0387*
S	0.7571	0.3193	2.3710	0.0258*

由表 6 可知，变量 M、I、S 是显著的，其 p 值小于 0.05。因此，使用 Lasso 回归方法建造的模型为：

$$Y = -0.48M - 0.4I + 0.76S$$

使用标准化后的数据进行计算，将系数还原为：

$$Y = 347.192 - 0.001730694M - 0.0005035332I + 0.1656068S$$

6. 主成分回归法

6.1. 主成分回归法介绍

主成分回归也是一种“降维”的方法。主成分回归可以使得原始数据信息损失最少，在这种前提下，通过线性变换将原始自变量的集合从高维空间映射到低维空间[9]。通过对原始自变量的线性组合组成新的主成分变量，利用新的主成分对数据进行普通最小二乘回归，然后再回到原来的自变量。主成分回归适用于数据存在多重共线性的情况。但主成分回归有时难以解释每个主成分的具体含义，有时需要对得到的主成分进行因子旋转，来解释得到的主成分。

主成分回归的步骤为：

1) 首先对数据进行标准化，得到标准化后的矩阵 X，及其对应的协方差阵 Z。

2) 求 Z 的前 m 个特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$，及对应的特征向量 $a_1, a_2, \cdots, a_m$，它们是标准正交的。

3) 求第 i 主成分 F_i，并且满足 $F_i = Xa_i, i = 1, 2, \cdots, m$。

4) 确认在第 n 个主成分时，前 n 个主成分可以完全解释要求的信息量。再用最小二乘法对前 n 个主成分进行回归。

6.2. 主成分回归法建模

进行主成分回归时，首先需要选择主成分。使用 validationplot 函数可视化 RMSE (均方根误差)，并利用交叉验证检验 RMSE。可以得到表 7、表 8，表 7 中展示了交叉检验的 RMSE，表 8 中为响应变量能被主成分解释方差的百分比。图 6 为可视化 RMSE。图 7 为碎石图。

Table 7. Cross-validated RMSE
表 7. 交叉验证的 RMSE

	intercept	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps
CV	46.41	45.64	49.85	43.93	46.64	41.04	42.41
adjCV	46.41	45.46	49.26	43.37	46.07	40.53	41.79

Table 8. Percentage of response variables that can be explained by the variance of the principal components
表 8. 响应变量能被主成分解释方差的百分比

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps
X	49.46	82.02	94.58	97.31	99.26	100
Y	5.129	18.7	39.02	40.66	47.77	48.89

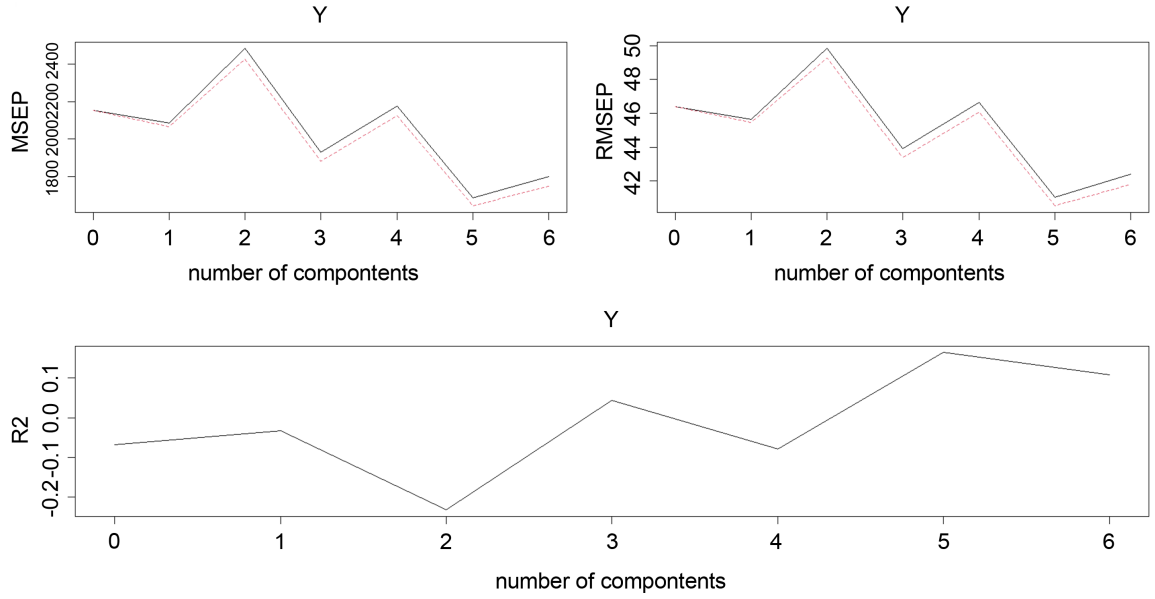

Figure 6. Visualization RMSE
图 6. 可视化 RMSE

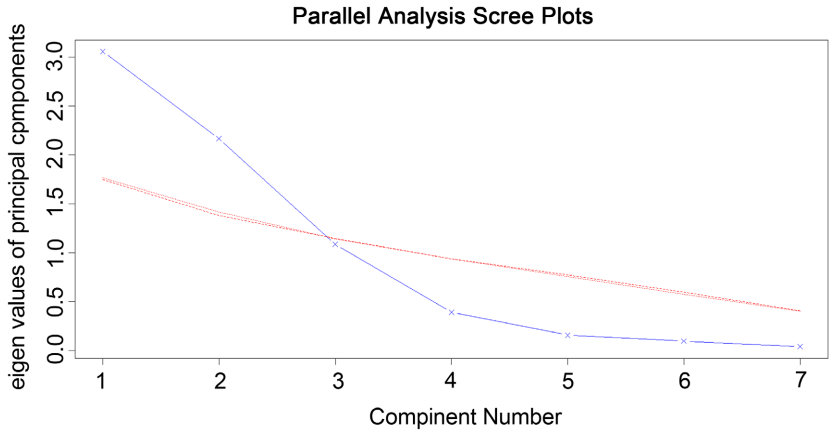

Figure 7. Gravel map
图 7. 碎石图

综合上述表格及图片，选择三个主成分较为合适。此时，误差较小，同时再增加主成分个数对方差的解释量不会增加太多。因此，选择三个主成分进行最小二乘回归。通过 princomp 函数得到，前三个主成分的累计贡献率为 94.6%。用 Y 对前三个主成分进行普通最小二乘回归，可以得到主成分回归的模型为：

$$Y = -0.13Z_1 + 0.25Z_2 - 0.51Z_3$$

其中

$$Z_1 = 0.40M + 0.40G + 0.42P + 0.45B + 0.12I + 0.54S$$

$$Z_2 = 0.48M + 0.48G - 0.41P - 0.38B - 0.48I$$

$$Z_3 = 0.26M + 0.27G - 0.35P + 0.80I - 0.30S$$

同时，主成分回归的描述性统计量见表 9。

Table 9. Descriptive statistics for model construction by principal component regression
表 9. 主成分回归法构建模型的描述性统计量

自变量	系数估计值	标准误差	t 值	p 值
z_1	−0.1293	0.0843	−1.5350	0.1361
z_2	0.2593	0.1039	2.4960	0.01872*
z_3	−0.5108	0.1672	−3.0540	0.00491**

Z_1 不显著，因此，最终主成分回归得到的模型为：

$$Y = -0.0126M - 0.0177G + 0.076P - 0.095B - 0.528I + 0.153S$$

由于数据进行了标准化，应将模型的系数进行还原。使用原始数据的标准差及均值等，将数据代入模型中，最终得到：

$$Y = 328.0003 + 0.001125289P - 0.03769546B - 0.0006646639I + 0.03333925S$$

7. 各种回归方法精度的比较

通过上述计算及建模，利用五种方法，将空气质量数据作为原始数据，我们得到了五个模型。将数据代入模型表达式，可以得到每种方法对应的预测结果，见附表 3。通过与真实值进行比较，可以得到表 10。表 10 中数据是通过计算得到模型的平均绝对误差与 RMSE (均方根误差)，可以对各个模型的精度进行比较。

Table 10. Comparison of the accuracy of each model
表 10. 各个模型精度的比较

	普通最小二乘	逐步回归	岭回归	Lasso 回归	主成分
平均绝对误差	40.2502	26.5213	33.6730	35.1313	28.4649
RMSE	47.5713	32.7027	41.9862	43.9819	36.7593

由表中可以得到，从平均绝对误差和 RMSE 两方面，逐步回归法与主成分回归法表现最好。普通最小二乘法表现最差，这也验证了最小二乘法在存在多重共线性及实际问题中，不如一些对最小二乘法进行改进的方法精度更高、适应性更强。岭回归和 Lasso 回归的表现比较接近。我们知道岭回归、Lasso 回归、主成分回归在应对多重共线性时，可以有效减弱多重共线性的影响。在本文结果中也得到体现。

8. 结论

通过对模型精度的比较，逐步回归方法得到的模型较为准确。我们先对逐步回归方法获得的模型进行分析。逐步回归法得到的模型为：

$$Y = 394.8733 - 0.001802806M - 0.008439665P - 0.0006168282I + 0.1503535S$$

结合自变量的含义，我们得到空气质量与人均可支配收入、人口数、工业产品产量呈负相关，与政府科学技术预算支出呈正相关。这与常识相吻合。尽管有些省份人均可支配收入较高，但是空气质量不是很好；相反，在一些经济欠发展的省份，空气质量较好。同样地，空气质量与工业产品产量息息相关，这是因为进行工业生产时，往往会产生大量的有毒有害气体，对空气造成污染，例如：炼钢等。

结合其他方法生成的模型，可以得到其他模型自变量系数的分析与上述逐步回归模型的分析基本相同。但是注意到在主成分回归模型与岭回归模型中，环境保护预算支出的系数为负值，这与常识不符。

我认为是本文中建模不够准确，所使用的数据量较小造成的失误，还可以进一步改善。

除此之外，综合来看所有模型，可以发现自变量 I 与自变量 S 基本出现在所有模型中，这说明工业产品产量与科学技术预算支出对空气质量的影响比较大。尤其是工业产品产量，这提示我们，要想改善环境质量、提高空气质量，进行产业结构的调整比较重要。应减小经济对第二产业的依赖，才能更好地做到经济与环境协调发展。同时，提高对科学技术预算的支出，加大对高新产业的支持，也能达到改善环境与空气质量的目的。

参考文献

[1] 黄煜宁, 尤学敏, 温芫姚, 等. 福州市居民人均可支配收入与空气质量因子的关系[J]. 福建林业, 2022(2): 37-40.

[2] 金仁浩, 曾国静, 赵欣然. 北京地区空气质量影响因素分析及预测研究[J]. 黑龙江科学, 2022, 13(8): 46-50.

[3] 刘亦文. 碳减排约束政策对中国城市空气质量的影响研究[J]. 湖南大学学报(社会科学版), 2022, 36(2): 73-81.

[4] 姚作芳, 刘兴土, 杨飞, 闫敏华. 几种方法在粮食总产量预测中的对比[J]. 干旱地区农业研究, 2010, 28(4): 264-268.

[5] Hoerl, A.E. and Kennard, R.W. (2000) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **42**, 80-86. https://doi.org/10.1080/00401706.2000.10485983

[6] 杨楠. 岭回归分析在解决多重共线性问题中的独特作用[J]. 统计与决策, 2004(3): 14-15.

[7] 高惠璇. 两个多重相关变量组的统计分析(3) (偏最小二乘回归与 PLS 过程) [J]. 数理统计与管理, 2002(3): 58-64.

[8] 郭文军. 中国区域碳排放权价格影响因素的研究——基于自适应 Lasso 方法[J]. 中国人口·资源与环境, 2015, 25(S1): 305-310.

[9] 付凌晖, 王惠文. 多项式回归的建模方法比较研究[J]. 数理统计与管理, 2004(1): 48-52.

附 录

Table S1. Raw data
附表 1. 原始数据

region	*Y*	*M*	*G*	*P*	*B*	*I*	*S*
北京	276	69433.5	164,889	2189	236.9	962.68	410.96
天津	245	43854.1	101,614	1387	60.72	15303.58	118.17
河北	205	27135.9	48,564	7464	509.27	117276.36	101.76
山西	224	25213.7	50,528	3490	260.28	146008.44	66.09
内蒙古	294	31497.3	72,062	2403	149.37	122535.93	32.38
辽宁	287	32738.3	58,872	4255	97.94	39674.47	72.71
吉林	305	25,751	50,800	2399	131.53	10414.82	39.94
黑龙江	303	24,902	42,635	3171	220.27	13245.58	42.98
上海	319	72232.4	155,763	2488	181.88	7498.13	406.2
江苏	304	43390.4	121,231	8477	336.9	63784.07	584.39
浙江	334	52397.4	100,620	6468	220.59	31370.8	472.13
安徽	311	28103.2	63,426	6105	190.83	44162.48	369.98
福建	364	37202.4	105,818	4161	156.42	27214.34	149.44
江西	335	28016.5	56,871	4519	218.27	21524.52	195.74
山东	223	32885.7	72,151	10,165	291.54	62687.42	298.62
河南	230	24810.1	55,435	9941	272.63	40744.86	254.28
湖北	309	27880.6	74,440	5745	219.18	34806.49	287.85
湖南	309	29379.9	62,900	6645	245.58	27286.08	220.66
广东	331	41028.6	88,210	12,624	517.76	43874.03	955.73
广西	357	24562.3	44,309	5019	100.74	28047.23	66.26
海南	361	27904.1	55,131	1012	57.84	2674.02	35.67
重庆	331	30823.9	78,170	3209	179.71	13992.19	82.87
四川	280	26522.1	58,126	8371	264.02	35541.48	181.7
贵州	362	21795.4	46,267	3858	146.15	28808.77	113.19
云南	366	23294.9	51,975	4722	163.97	34868.89	64.94
西藏	366	21744.1	52,345	366	48.93	1097.57	8.99
陕西	250	26,226	66,292	3955	190.34	88627.85	56.45
甘肃	312	20335.1	35,995	2501	114.03	13673.37	32.07
青海	337	24037.4	50,819	593	73.51	4616.26	10.56
宁夏	301	25734.9	54,528	72	49.48	13682.4	27.91
新疆	279	23844.7	53,593	2590	82.59	40279.69	41.25

Table S2. Standardized data
附表 2. 标准化后的数据

	Y	*M*	*G*	*P*	*B*	*I*	*S*
1	−0.60346	2.949772	3.004045	−0.75859	0.379847	−1.01979	1.068273
2	−1.282529	0.929445	0.984112	−1.01872	−1.1515	−0.62434	−0.32929
3	−2.158747	−0.391	−0.70941	0.952309	2.747276	2.187596	−0.40762
4	−1.742544	−0.54282	−0.64671	−0.33662	0.583065	2.979892	−0.57789
5	−0.209162	−0.04653	0.04072	−0.68918	−0.38096	2.33263	−0.73879
6	−0.3625	0.05149	−0.38035	−0.0885	−0.82799	0.0477	−0.54629
7	0.0317982	−0.50039	−0.63803	−0.69048	−0.53602	−0.75914	−0.70271
8	−0.012013	−0.56744	−0.89868	−0.44009	0.2353	−0.68109	−0.6882
9	0.3384746	3.170836	2.712715	−0.66161	−0.09838	−0.83957	1.045553
10	0.0098928	0.89282	1.610347	1.280867	1.249043	0.712529	1.896103
11	0.6670564	1.60421	0.95238	0.629264	0.238081	−0.18128	1.360255
12	0.16323	−0.3146	−0.23497	0.511529	−0.02059	0.171458	0.872665
13	1.32422	0.404076	1.118316	−0.11899	−0.31968	−0.29589	−0.18003
14	0.6889619	−0.32145	−0.44422	−0.00288	0.217916	−0.45279	0.040969
15	−1.764449	0.063132	0.043562	1.828355	0.854776	0.682289	0.532043
16	−1.611111	−0.5747	−0.49006	1.755703	0.690411	0.077216	0.320396
17	0.1194201	−0.33218	0.116634	0.394766	0.225826	−0.08654	0.480635
18	0.1194201	−0.21377	−0.25176	0.686673	0.455293	−0.29391	0.159919
19	0.6013401	0.706279	0.556215	2.625911	2.821071	0.163504	3.668612
20	1.1708819	−0.59427	−0.84524	0.159294	−0.80365	−0.27293	−0.57708
21	1.2585037	−0.33033	−0.49977	−1.14034	−1.17654	−0.9726	−0.72309
22	0.6013401	−0.09971	0.235707	−0.42776	−0.11725	−0.6605	−0.49779
23	−0.515838	−0.43948	−0.40416	1.246486	0.615573	−0.06627	−0.02605
24	1.2804091	−0.81281	−0.78274	−0.21727	−0.40895	−0.25193	−0.35307
25	1.3680309	−0.69437	−0.60052	0.062964	−0.25406	−0.08482	−0.58338
26	1.3680309	−0.81686	−0.58871	−1.34987	−1.25398	−1.01607	−0.85044
27	−1.173002	−0.46287	−0.14348	−0.18581	−0.02485	1.397604	−0.6239
28	0.1851364	−0.92815	−1.11065	−0.6574	−0.68813	−0.66929	−0.74027
29	0.7327728	−0.63573	−0.63742	−1.27624	−1.04033	−0.91904	−0.84295
30	−0.055824	−0.50166	−0.51902	−1.44522	−1.2492	−0.66904	−0.76013
31	−0.537744	−0.65095	−0.54887	−0.62853	−0.96141	0.064389	−0.69646

Table S3. The true value and the predicted value of each model
附表 3. 真实值与各模型的预测值

Y	最小二乘法	逐步回归法	岭回归	Lasso 回归	主成分回归
276	292.6335	312.6951	307.2081	296.7287	334.4882
245	285.4630	312.8639	309.9671	284.5572	320.4594
205	234.4766	228.2598	289.0943	259.2742	238.3064
224	220.1106	242.5009	292.1471	242.2690	222.0216
294	231.8468	249.3381	297.1340	237.7408	240.3155
287	273.2776	287.3260	306.6442	283.7404	303.6376
305	287.9074	328.1240	308.7707	304.8220	319.7201
303	286.4920	321.9258	306.3756	305.3533	315.3435
319	289.3657	300.5088	307.8640	287.9146	332.1580
304	261.2228	295.1272	298.5460	338.3113	299.4352
334	277.4294	298.3654	304.5555	320.6157	320.5612
311	271.0336	322.1160	303.9846	338.6040	308.9281
364	279.5076	299.0888	306.4984	295.0880	312.6067
335	282.3525	322.9890	305.5953	321.2162	316.1934
223	261.7711	257.5632	299.7352	309.3941	294.2536
230	272.7424	280.4821	302.3795	326.7512	308.6034
309	275.7116	318.8102	304.2455	330.0597	311.2776
309	279.4718	302.9650	304.3692	320.1444	314.2992
331	271.1778	332.3020	296.2325	413.7753	323.4811
357	279.0912	301.6321	307.7403	302.3851	312.2967
361	291.7778	339.9019	311.2986	304.3250	326.2489
331	286.1187	316.4947	307.2663	301.5185	317.7202
280	275.3441	282.8017	303.1048	314.4232	308.4264
362	278.7104	322.9616	306.5834	314.4797	310.3373
366	275.6804	302.1168	305.5532	300.9103	304.7671
366	292.5660	353.3522	311.6683	311.1670	326.0893
250	248.8009	269.7522	299.5497	267.6420	265.0153
312	286.2781	333.8778	308.8613	311.0964	317.9490
337	290.8067	345.4422	310.7314	305.7690	323.0015
301	286.2736	343.9311	310.3967	301.2234	317.5658
279	272.9750	312.2289	306.9490	293.3472	300.9135